Group #8
Members: Vaishnavi Borwankar, Niharika Chunduru, Kyle Zhao, Peter Kryspin
Date of Submission: 5/7/23
Link to the app: https://niharika-chunduru.shinyapps.io/stat436-group8-project/
Link to code: https://github.com/NiharikaCNR/stat436-group8-project/blob/main/Milestone_3.R

STAT 436 - Milestone 3

Introduction:

Air travel is essential, but flights are almost synonymous with delays themselves. Thus, it is crucial not only for airline companies to minimize the frequency of delays to maximize profits and customer loyalty, but also for consumers to book flights without delays. Through the use of several interactive visualizations, we provide an interface that can be used to examine and interpret the trends of delayed flights. With our interface, airlines will be able to determine which flight lengths for which days are associated with higher frequencies of delays. Furthermore, airports will be able to compare how often and on which days flights departing from or arriving there are delayed. With the interface, consumers may also compile information about airports and airlines to plan a trip with the least likelihood of a delay.

Literature Review:

Many studies have been conducted in relation to airplane delays. For instance, a study (Brueckner) utilizes regression techniques on variables they generate based on delay times in addition to weather factors. This study goes one step further than our analysis by examining how an initial delay in an airplane's departure can propagate through an airline's flights and cause further delays. Their findings include greater delay times for airlines for which an airport is not their hub, greater delay times later in the day from propagation, and greater costs due to delays for airlines which have quicker turnaround times. The multitude of variables they analyzed impressed upon us the many layers of analysis required for flight delays, which inspired us to create an interface that allows a user to investigate many variables of the dataset and understand possible associations.

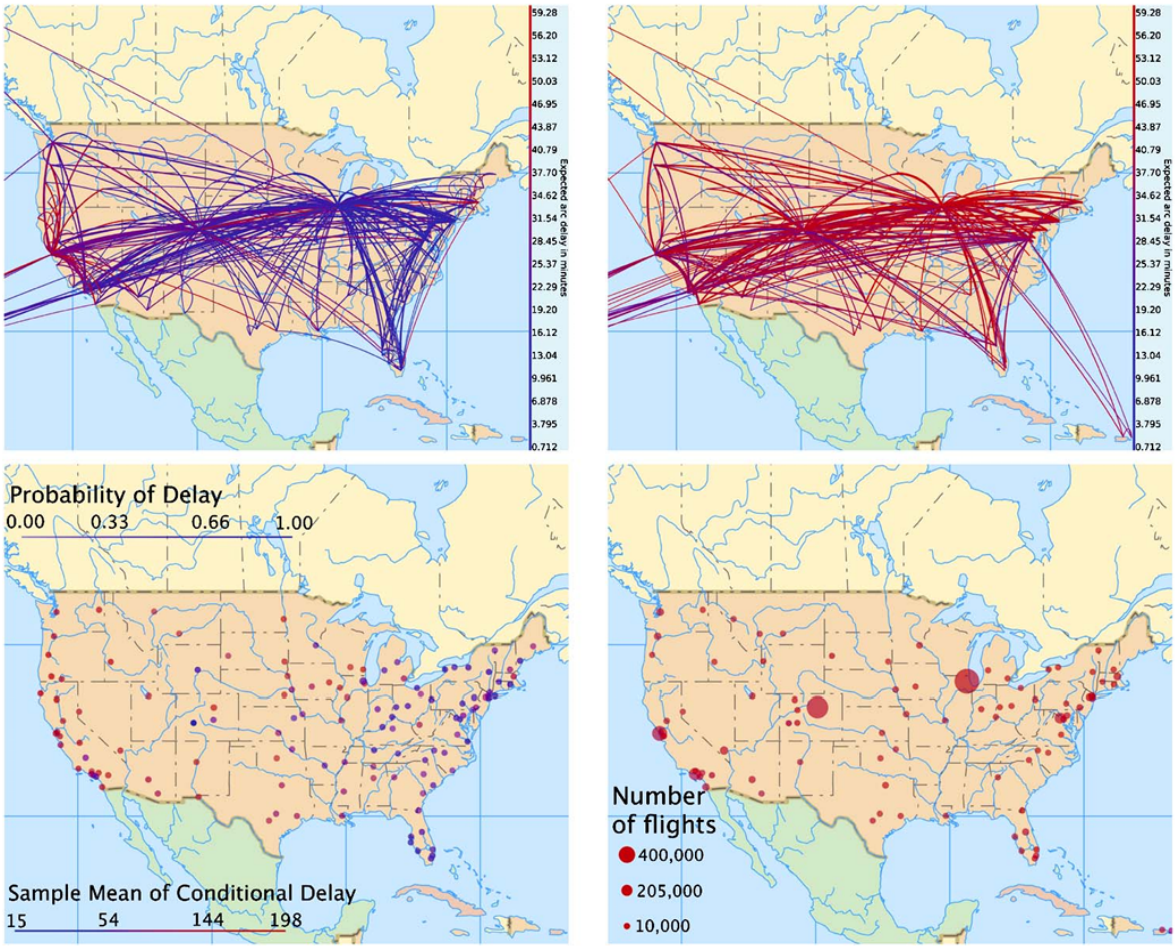

Figure 1. United Airlines in 1987 and 2008.

The visualization above (Figure 1.) shows United Airlines flights from 1987 (left two charts) compared to 2008 (right two charts). The color represents the average delay time, the arcs represent each flight and the diameter of the circles represent the number of flights at each airport United Airlines serviced (Dey). We drew direct inspiration from the bottom two plots shown and combined them into the first visualization in our interface: the size of each point represents the log of the total number of flights, while the color of each point represents the proportion of flights delayed.

Another paper, entitled "The Economic Cost of Airline Flight Delays" by Borenstein and Rose (2012), estimates the economic cost of airline delays in the United States. The paper discusses the factors that contribute to airline delays, such as weather, airport congestion, and aircraft maintenance issues. The authors note that while some delays are within the control of airlines, such as maintenance issues and crew scheduling, others, such as weather-related delays, are beyond their control.

The authors estimate the direct costs of delays to airlines and passengers, such as additional fuel costs and lost productivity, as well as the indirect costs to the wider economy, such as lost business opportunities and reduced consumer welfare. They also consider the

benefits of delay reduction, such as improved airline competitiveness and increased consumer welfare.

The authors find that delays cost the US economy $32.9 billion in 2007, with the majority of these costs being borne by passengers and airlines. They note that reducing delays by even a small amount could have significant economic benefits, such as increased productivity and consumer welfare, which inspired the premise of our project.

Cost component	Cost (in billions)
Cost to airlines	$ 8.3
Cost to passengers	$16.7
Cost from lost demand	$ 3.9
Total direct cost	$28.9
Impact on GDP	$ 4.0
Total cost	$32.9

Table 1. High cost of Inefficiency

The table above (Table 1.) shows the data summarized by the Total Delay Impact Study from 2007 and the breakdown of the total cost, determined with novel accounting methods (Guy).

Designs:

To translate abstract concepts from the literature review to resolve specific problems identified in the introduction, it is important to first identify the key findings and insights from the studies and understand how they relate to our problems. From the Borenstein and Rose study, we find that delays have significant economic costs and from the Brueckner study, that delays can propagate through airlines and cause a chain-reaction, leading to higher costs for airlines. This highlights the importance of analyzing which airlines and airports tend to have more delays so that they can improve their operations.

Interface

The application aims to provide four different visualizations of the data that each inform a user's understanding in a different way. The first visualization is a **leaflet map** that visualizes airports across the United States, where each point encodes an airport, the size of the point encodes the log of the total flights, and the color encodes the proportion of flights delayed. The interactivity for this visualization stems from clicking on and moving the leaflet map itself as well as the ability to switch between the views of origin and destination airports. The intention behind displaying this map on top is to give the user an overview of what airports are in their

area as well as how they may compare to each other. For instance, near Madison, WI the user can find that there are only two airports – the Dane County Airport and Mitchell International Airport in Milwaukee. Since the proportion of delayed flights is similar, if we consider them origin airports, the user may want to investigate these two airports further. Next to the leaflet map is a brief introduction to the application as well as a few suggested ways to use it.

The **Distribution of Flight Lengths** visualization allows a user to view a density plot of flights that were delayed or on time, faceted by different days of the week, for any number of selected airlines using a checkbox input. The advantage of this visualization is that it allows a user to view how the distribution of flights changes across the week for certain airline(s). If a certain day's distribution of delayed flights differs significantly from another, they may want to investigate why such a difference in the distribution is present, or avoid the day for which a chosen flight length has a higher chance of a delay. For instance, if a user wants to book a flight that will be around 2 hours long with American Airlines, they can find that the difference between the probability of a 'delayed' flight and an 'on-time' flight is the greatest on Wednesdays and is the smallest on Sundays. Furthermore, the check-box input filters the **data table** below this visualization. This allows a user to scroll through previous flights of the selected airline(s). This is especially useful if they use the built-in search box to filter for a certain day of the week or a given airport. Using the ongoing example, they could type "MKE" into the search box to look into which American Airlines flights from Mitchell International Airport had a delay history to be better informed on their flight bookings.

The **Flight Traffic per Airport** visualization allows a user to visualize how the number of flights across a given week varies for any combination of airports. A user can interact with this visualization by searching for the flight traffic of any airport(s) over a given week, allowing them to determine which days may be busier to increase their chances of booking a flight. From this view, we may easily find the overall trend that many airports tend to have a massive spike in flight traffic in the middle of the work week, Wednesday through Friday, decrease on Saturday, then increase slightly from Sunday through Tuesday. Furthermore, using the previous Dane County Airport versus Mitchell International Airport example, a user can see that the Mitchell International Airport vastly outnumbers the number of flights of Dane County Airport for every day of the week even though they have similar trends in number of flights across a week, which may cause a user to prefer flights from Mitchell International Airport due a greater number of flights while having a similar distribution to Dane County.

Synthesis

One domain-specific problem may be if someone wanted to determine when on a daily basis nearby airports are the least likely to experience delays. For this task, a line plot with the proportions of delays for an airport on the y-axis versus the day of the week on the x-axis may prove useful. Furthermore, a user could brush over a static map to select airports in the vicinity, displaying each selected airport as a separate line on the plot for the purposes of comparison.

Another potential domain-specific problem may be determining at what times of the day delays for a given airline tend to occur. Airlines could be selected using a checkbox input, and for each hour in the day, the proportion of flights delayed are graphed using a line plot, with different lines indicating different airlines. This would allow a user to directly compare at what times certain airlines perform better than others or not, which would better inform the purchase of their tickets.

Conclusion:

In summary, predicting when airplane delays will happen is a multi-faceted, difficult problem for which we hope our application interface will help. Oftentimes, shorter flights tend to have a lesser chance of being delayed, while the proportion of flights delayed for airports tends to be higher if we consider them arrival airports rather than departure airports.

While we believe our interface provides a useful tool for consumers, airlines, and airports through visualization of many previous flights, further statistical techniques may be useful to determine a comprehensive model for predicting delays. For instance, machine learning or deep learning methods such as a random forest may provide a way for a user to more accurately predict based on a given day, airline, and airport combination whether the flight will be delayed or not. These predictive models, together with the interface we created, may better inform everyone involved in the air travel industry to lessen the impact of flight delays. There are some limitations to the visualization, including the fact that it only includes data from the United States and does not take into account other factors that may contribute to delays. As next steps, we consider expanding the visualization to include data from other countries and incorporating additional variables such as weather patterns and air traffic control data.

Works Cited:

Borenstein, S., & Rose, N. L. (2012). The economic cost of airline flight delays. The Journal of Transport Economics and Policy, 46(2), 1-27.

Brueckner, Jan K, et al. "Airline Delay Propagation: A Simple Method for Measuring Its Extent and Determinants." Transportation Research Part B: Methodological, Pergamon, 30 May 2022, https://www.sciencedirect.com/science/article/pii/S0191261522000741.

Dey, Tanujit, et al. "A Graphical Tool to Visualize Predicted Minimum Delay Flights." Journal of Computational and Graphical Statistics, vol. 20, no. 2, 2011, pp. 294–297., https://doi.org/10.1198/jcgs.2011.5de..

Guy, Ann Brody. "Flight Delays Cost More than Just Time." Berkeley Engineering, 4 Nov. 2010, https://engineering.berkeley.edu/news/2010/11/flight-delays-cost-more-than-just-time/.